Integrating Medical Scientific Knowledge with the Semantically Quantified Self

Allan Third[1], George Gkotsis[2], Eleni Kaldoudi[3], George Drosatos[3], Nick Portokallidis[3], Stefanos Roumeliotis[3], Kalliopi Pafili[3], and John Domingue[1]

[1]Knowledge Media Institute, Open University, Milton Keynes, UK
{allan.third,john.domingue}@open.ac.uk

[2]King's College London, Biomedical Research Centre Nucleus, London, UK
george.gkotsis@kcl.ac.uk

[3]School of Medicine, Democritus University of Thrace, Alexandroupolis, Greece
{kaldoudi@med,gdrosato@ee}.duth.gr,portokallidis@gmail.com,
{st_roumeliotis,kpafili}@hotmail.com

Abstract. The assessment of risk in medicine is a crucial task, and depends on scientific knowledge derived by systematic clinical studies on factors affecting health, as well as on particular knowledge about the current status of a particular patient. Existing non-semantic risk prediction tools are typically based on hardcoded scientific knowledge, and only cover a very limited range of patient states. This makes them rapidly out of date, and limited in application, particularly for patients with multiple co-occurring conditions. In this work we propose an integration of Semantic Web and Quantified Self technologies to create a framework for calculating clinical risk predictions for patients based on self-gathered biometric data. This framework relies on generic, reusable ontologies for representing clinical risk, and sensor readings, and reasoning to support the integration of data represented according to these ontologies. The implemented framework shows a wide range of advantages over existing risk calculation.

Keywords: Health, Comorbidities, Risk factor, Scientific modelling, Knowledge capture, Semantics, Ontology, Linked Data.

1 Introduction

An important task in medicine is the assessment of risk. This depends on scientific knowledge derived by rigorous clinical studies regarding the (quantified) factors affecting clinical changes. Existing risk prediction tools typically only cover a very limited range of patient states, and the scientific knowledge informing the predictions is hardcoded into the tool. This makes them limited in application, particularly for patients with comorbidities (multiple co-occurring conditions), and rapidly out of date. An explicit representation of this knowledge, covering a wide (and, more im-

portantly, expandable) range of risks and outcomes, would enable more sophisticated and maintainable risk prediction, prevention and management.

In order actually to assess risk for an individual patient, it is necessary to link this generic clinical knowledge of risk to actual data relating to that patient's physical state. Traditionally, a doctor will make specific observations of a patient, and mentally determine the relevant known clinical evidence to make a risk prediction. In recent years, risk calculators based on individual clinical studies have been implemented as, e.g., web tools, where a patient can enter certain observations and be presented with numerical risks. With the advent of "Quantified Self" (QS) devices for low-cost and easy collection of individual physical and emotional data, there is a significant opportunity for personalized predictive medicine to combine this data with up-to-date knowledge of risk.

We present here a framework for calculating clinical risk predictions for patients based on self-gathered biometric data, using Semantic Web technologies at the core. The framework is shown to enable a large body of medical knowledge to be encoded in a common framework, and faithfully applied to QS data to perform automatic risk calculation, providing qualitative and quantitative improvements over the state of the art.

2 Previous Work

Existing algorithms for risk prediction for, e.g., cardiovascular risk, include the Framingham equation [1], the Joint British Societies (JBS) formula [2] and the ASSIGN score [3]. These take account of a limited set of risk factors and possible outcomes, as these have been produced by specific clinical studies – thus can be limited in application. For example, the ASSIGN score is specialized for Scottish populations, and, while Framingham includes diabetes as a risk factor, it is omitted from the JBS formula (diabetic patients are always high-risk). The Framingham equation takes account of 9 different patient observables and predicts the risk of only one outcome. More fundamentally, each of these hardcode the scientific knowledge about risk into the prediction formula itself, thus requiring new versions to be created to accommodate new scientific knowledge. This limited and non-extensible approach motivates our construction of a generic semantic model.

As there is no other model addressing the concept of risk factor, to the best of our knowledge, we compare related work addressing similar concepts and level of abstraction. A number of models have been proposed for capturing various aspects of clinical research at various levels of granularity. In particular, the Ontology-Based eXtensible data model (OBX) [4] has been developed to represent results of clinical research in order to promote data reuse, but does not address the concept of population-level risk factor. Models maintained by the Clinical Data Interchange Standards Consortium (CDISC) [5], and the Ontology of Clinical Research (OCRe) [6] take a more top-down approach to the modelling of clinical research and focus on data interchange formats and on the conceptual modelling of proposed and ongoing clinical trials. Overall, existing models aim to support the process of generating new scientific

knowledge in medicine, rather than represent the actual knowledge itself, which is required for the task of risk prediction. In order to capture this scientific knowledge, we have developed an ontology for medical risk factors.

The "Quantified Self" (QS) refers to the use of technology for automated tracking of various measurements related to oneself (e.g., daily step count, distance walked, weight, and so on). Although considered a new trend with the potential to transform healthcare, it has received only a small amount of attention from the Semantic Web community. The MoodMap app [8] represents emotional states using an ontology, in order to support analysis of mood as tracked in the workplace, but does not concern itself with other Quantified Self measurements. An ontology for QS was presented in [9], but this is very high-level and at an early stage, lacking the detail needed to implement a Semantic Web system making use of it. This paper presents a detailed and practical ontology for representing QS measurements semantically, in a way which encourages flexibility and reuse, linking to other concepts related to each measurement, and which is usable in practical systems.

Finally, in order to achieve the integration between medical knowledge and QS data, it is necessary to express rules describing when a particular piece of knowledge is relevant to an individual on the basis of gathered data. There are two main candidate standards for representing rules for the Semantic Web – SWRL [10] and RIF [11] - as well as widely-used systems such as Jena [12]. Unfortunately, none of these rule systems can offer the expressive power needed to describe the conditions necessary to personalize a risk factor description to an individual person's data. In particular, it is common that the conditions under which clinical risks can be identified depend on a range of functions, e.g., body mass index, or the time since the occurrence of a myocardial infarction. This requirement rules out Jena, SWRL, and the Core dialect of RIF. These conditions can often also require disjunction to express correctly – "if estimated glomerular filtration rate is less than 44 OR chronic kidney disease is diagnosed at stage 3 or 4 or 5", and negation ("if the patient is male and does not have a family history of ischemic heart disease"). While the RIF Basic Logic Dialect (RIF-BLD) does support disjunction (where SWRL and Jena do not), and is compatible with OWL [13], it does not support negation. It is therefore necessary to develop a dedicated rules expression format, in a way which is, by design, easily interpreted and evaluated, supports the required logical features, and which allows the contents of rules to be easily authored and understood by clinicians.

The risk ontology implements the model described in [7]. The measurements ontology and data integration via rules are presented for the first time here.

3 Risk Factors

In medicine, risk is the probability of a negative outcome on the health of a population of subjects. The agents responsible for that risk are called risk factors when they aggravate a situation and are used to predict (up to a degree) the occurrence of a condition or deterioration of a patient's health dividing the population into high and low

risk groups [14]. The following paragraphs present our model of the concept of **risk factor** in medicine [7] which is shown schematically in **Fig. 1**.

In general, risk factors can be: environmental (e.g. chemical, physical, mechanical, biological and psychosocial elements that constitute risk factors to public health); demographic (e.g. age, sex, race, location, occupation); genetic; behavioral and life-style related (e.g. smoking, overeating, unprotected sexual life, excessive alcohol drinking, drug abuse and sedentary lifestyle); and biomedical (i.e. conditions present in a patient that can influence his/her health by creating or affecting other conditions). Extending work on general risk analysis [15,16], we can present a risk factor as a triplet, which includes the **source** of the risk, the outcome (**target**) and an expression of their association. The source of the risk is an agent (an event, a condition, a disorder or any other factor) that is shown via empirical studies to be associated with a consequence, that is, the outcome. The outcome itself is a negative health condition or disorder. Most often the outcome itself is found to be a source of another risk factor.

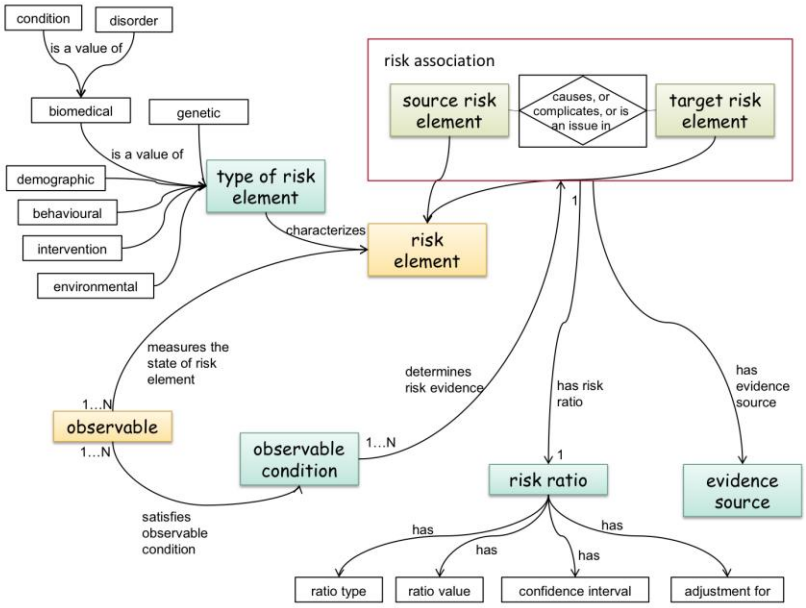

Fig. 1. Basic concepts and their relationships.

Thus in the general case the source and the outcome can both be treated as health related conditions (including disorders). In this work, we collectively refer to both the source and the outcome as **risk elements**. A **risk association** between the source and the outcome is a complex construct which describes the type of relation, the likelihood of an outcome to occur, and the initial conditions under which such likelihood can be estimated. The existence of a risk factor is not a determinant of consequence but the degree of its influence can be statistically calculated. The way to measure the likelihood requires a certain quantitative biomarker and observational studies that

statistically calculate a probability. Different study designs and analyses can generate different types of probability measures [17] - a **Risk Ratio** (RR), such as the Relative Risk or Hazard Ratio (HR). A probability determined from a clinical study lies within a confidence interval, and the study design/analysis may have been adjusted, or not, for certain factors (for example, age, sex, and so on). In order to be able properly to represent risk factors, these must be included – especially where the goal is to produce personalized risk calculations.

An event, a condition, a disorder or any other factor becomes a risk source when certain conditions are met. These conditions are associated with one or more **observables**, which is either environmental or a physical or mental property of the patient. Therefore, in order to describe properly a risk association we have to state a specific observable that provides a measure/description of the risk source and the specific condition or value of this observable. For the same risk factor, a number of different risk associations can be measured in the literature, each association corresponding to a different observable or a different **observable condition** or even different combinations of observables corresponding to different concurrent risk sources. The circumstances under which a risk association is relevant to an individual are ascertained via an explicit logical expression that involves observables; this logical expression is termed 'observable condition'.

Finally, risk associations in medicine are determined from clinical studies as reported in evidence based medical literature. Thus, each association is directly related to an **evidence source** which is a specific scientific publication.

To ensure that the model can be seamlessly integrated into existing medical information systems, we adopt commonly used standards and controlled vocabularies in the description of the concepts presented above. For example, risk elements of type biomedical include an ICD-10 [18] classifier, of type demographic, a SNOMED-CT [19] classifier. Other controlled vocabularies used for risk elements of type environmental or intervention include SNOMED-CT, RxNorm [20], and EnvO [21]. Measurements and units follow the QUDT [22] and UO [23] ontologies. Evidence sources are described using their DOI and/or their PubMed identifier, while evidence level follows the OCEBM system [24]. In general, where available UMLS [25] codes are also used.

View	Risk factor	Observable condition ▲1	Ratio value
	central obesity [is an issue in] acute myocardial infarction	(waist circumference < 102 AND waist circumference ≥ 94) AND sex = 'male'	1.1
	central obesity [is an issue in] acute myocardial infarction	(waist circumference < 88 AND waist circumference ≥ 80) AND sex = 'female'	1.5
	central obesity [is an issue in] acute myocardial infarction	waist circumference ≥ 102 AND sex = 'male'	2.8
	central obesity [is an issue in] acute myocardial infarction	waist circumference ≥ 88 AND sex = 'female'	1.4

Fig. 2. Example of risk associations and corresponding observable conditions

Fig. 2 shows the risk associations relevant to the risk factor "central obesity is an issue in acute myocardial infarction", with the risk ratio values associated with patients who satisfy each observable condition, respectively. Although omitted here for

reasons of space, each of these ratios is also associated with the original publication providing the evidence for it, as well as a confidence interval and specific ratio type.

4 Measurements and Sensors

The aim of the readings and measurements ontology is to represent the concepts involved in the gathering of data from personal Quantified Self sensors. In particular, it is important to represent details which are common to measurements generically, while allowing details relevant to specific measurement types to be captured also. Crucially for data integration, each measurement should be associated with a canonical type (representing, e.g., "systolic blood pressure") and a unit (e.g., "mmHg"), both preferably denoted by terms in standardized external vocabularies where possible. **Fig. 3** illustrates the ontology for the CARRE measurements and sensors.

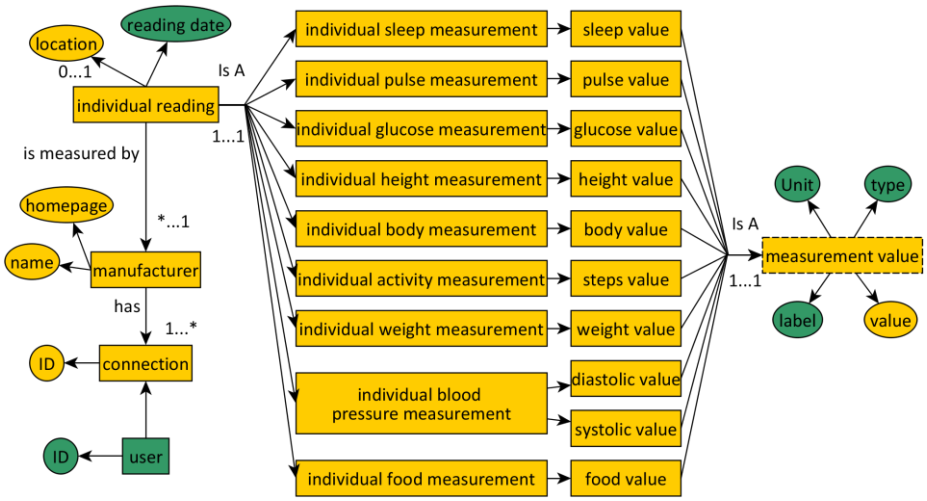

Fig. 3. The CARRE Sensors and Measurements Ontology, including some specific types of measurement to illustrate types of data which can be represented.

A **user** (an individual whose data is being represented using the terms of this ontology) has an identifier and **connections**. A connection represents that user's login details to the cloud data source, usually in practice provided by a device **manufacturer**, which has a name and a website.

Procedurally, data for an individual user is gathered from a manufacturer by means of the connection. Data is in the form of one or more device **readings**. Every device reading must of course have a date at which the reading was taken. Some manufacturers also provide location information in the form of latitude and longitude. A device reading may represent a set of **measurements**, all of which are semantically related. For example, a device reading may originate with the user stepping onto a set of body analysis scales, which can provide measurements of weight, body fat percentage, muscle mass, and so on. A reading may also have a provenance, which at the time of

writing is simply whether the measurement came from a device automatically, or was manually entered into a web form by the user, and an actuality: manufacturers may provide actual measurements from devices or users, or goal measurements (e.g., a target weight). Finally, a device reading may be associated with a textual note added by the user.

The device reading class may be sub-classed, for measurements of, for example, activity, weight, blood pressure, and so on. Each of these has properties relating to the type of measurement value represented: for example, an individual blood pressure measurement relates to both systolic and diastolic blood pressure valuess.

Every measurement value has a common structure. A measurement value has a measurement type and a unit, which are its type and unit expressed in an external vocabulary wherever possible, a value which can be an integer, string, floating point value, and so on, and a label, which is a human-readable string.

To ensure that the model can be seamlessly integrated into existing medical information systems, we adopt the commonly used standards and controlled vocabularies in the description of the concepts presented above. The FOAF ontology [28] is extremely widely used and well-known, and allows easy representation of data relating to people. Types of measurement are indicated with respect to the Logical Observation Identifier Names and Codes ontology (LOINC) [29] and the Clinical Measurements Ontology (CMO) [30], with preference given to CMO on the basis of coverage for the set of measurement types currently being used. For units, we use QUDT [22] and the Unit Ontology [23], with preference for the Unit Ontology, again, on the basis of coverage.

5 Data Aggregation and Enrichment

These two ontologies are generic, describing the structure of data relating to measurements and to risk factors. To be useful, we need to populate them with instances of particular measurements and risk factors, respectively. The output of the relevant data aggregation processes is Linked Data, expressed according to the vocabularies defined by the relevant ontology, and stored in an RDF quad-store (Virtuoso, [31]).

Measurement data is subject to some extra constraints compared to the risk factor data. While clinical knowledge relating to risk is generic, and therefore can (and, we would argue, should) be public, measurement data is specific to an individual, and, as personal health-related data, required to be kept private. We thus maintain a separation between them at the quad-store level. Risk data is stored in a (curated, for quality and safety purposes) publicly accessible RDF graph, where measurement data relating to an individual is stored in an authentication-protected RDF graph belonging to that individual, accessible only via HTTPS.

There is a wide range of different wearable and personal sensors available which can, usually via a smartphone connection, automatically upload measurements to a manufacturer service. Such devices exist to measure activity levels (step counts, distance travelled), heart rate, blood pressure, blood oxygen saturation, weight, body fat, and others. In this work we have developed aggregators for data from devices from

multiple manufacturers, including Fitbit, Medisana, iHealth, and Withings [32,33,34,35]. In each of these cases, the measurements are available for programmatic access via a Web API secured by some variant of the OAuth authentication schemes. Each such API is supported in the aggregator by a plugin module, which, when supplied with access tokens for a particular user, retrieves that user's measurements, enriches them with RDF, and stores them in the relevant graph in the quadstore. Once set up, measurements are retrieved automatically, according to either the device's or user's chosen sampling interval, unless the user chooses to revoke access.

The risk ontology was populated with scientific information on medical risk factors in the area of cardiorenal disease. Chronic cardiorenal disease is the condition characterized by simultaneous kidney and heart disease while the primarily failing organ may be either the heart or the kidney. The cardio-renal patient (or the person at risk of this condition) presents an interesting case example for exploring risk factors, as (a) it is a complex comorbid condition which involves and is affected by a number of related health disorders as well as lifestyle related factors; (b) chronic cardiorenal disease has an increasing incidence and a number of serious (and of increasing incidence) comorbidities, including diabetes and hypertension, and may lead to serious chronic conditions such as nephrogenic anemia, renal osteodystrophy, peripheral neuropathy, malnutrition, and various systemic diseases (e.g. rheumatoid arthritis, lupus erythematosus); and (c) prevention is of major importance. Good appreciation of risks therefore plays an important role for the various stages of cardiorenal disease evolution, from normal health condition, to chronic disease, to end-stage renal and/or heart failure.

The process of collecting risk factor data begins with a literature review by the medical experts, to identify risk associations and associated entities and properties according to the ontology model. Identified risk factors are recorded in a tabular format, which mirrors the structure of the model, and these are reviewed by multiple clinicians. Observables, evidence sources, risk elements and associations are then translated to RDF.

6 Data Integration

The integration between the medical scientific knowledge and the semantic QS data is achieved using observable conditions. Each specific risk association is associated with a list of relevant observables, and an observable condition written in terms of these observables. Observable conditions are built using two basic types of operators, logical and comparison operators. More precisely, we follow prefix notation syntax for logical operators and infix notation syntax for comparison operators, and support as logical operators the disjunction "OR" and the conjunction "AND", and as comparison operators the equality "=", inequality "!=", greater than ">", greater than or equal to ">=", less than "<" and less than or equal to "<=". We have also identified functions which occur in the current domain of application, and use the idea of "calculated observable" to represent them. For example, "time since myocardial infarction" is calculable given the current date and a (non-calculated) observation of a myocardial

infarction event. Generic functions such as averages over time are also important to take account of possible differences in sampling interval in measurement data. These calculated observables can be used in observable conditions.

Fig. 4 shows a user-friendly interface that is used to build the observable conditions. This interface is implemented in HTML5, CSS and JavaScript using the AngularJS framework. As output of this expression builder, we support two different formats, an abstract syntax tree format (**Fig. 5**.a) and a simple free text format (**Fig. 5**.b). The first one is more suitable for expression editors and other parsers because it follows formal JSON syntax, and the second is more suitable for humans and evaluation algorithms and tools because it follows formal plain text syntax.

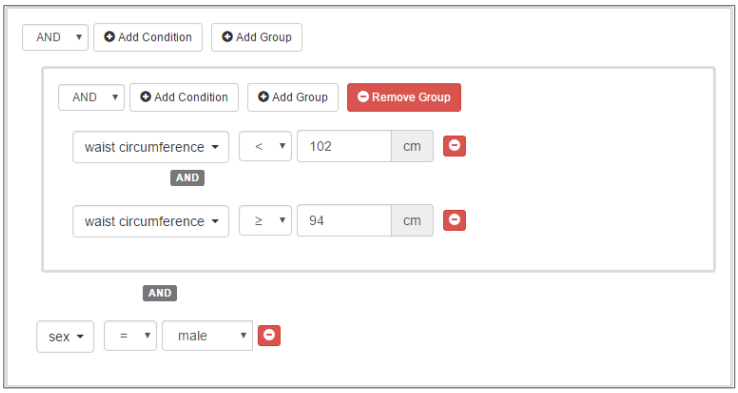

Fig. 4. Web based interface of expression builder.

```
{
    group: {
        operator: "AND",
        rules: [
            {
                group: {
                    operator: "AND",
                    rules: [
                        {
                            condition: "<",
                            field: "OB_80",
                            data: 102
                        },
                        {
                            condition: ">=",
                            field: "OB_80",
                            data: 94
                        }
                    ]
                }
            },
            {
                condition: "=",
                field: "OB_64",
                data: "male"
            }
        ]
    }
}
```

(a)

```
( OB_80 >= 94 AND OB_80 < 102 )
      AND OB_64 = 'male'
```

(b)

Fig. 5. Observable condition: (a) abstract syntax tree format (b) simple free text format.

The software evaluates these conditions by retrieving the relevant measurement data for the patient in question, and substituting values into the condition expression. The (boolean) result of this evaluation determines whether or not the condition's risk factor applies to that patient, and hence with what particular ratio the patient is at risk of its target.

For example, if we evaluate the expression of **Fig. 5** with observable values **waist circumference** (OB_80) equal to **98** and **sex** (OB_64) "**male**", the expression evaluates to *true*. Referring back to **Fig. 2**, we can see that this therefore means that with regard to the risk factor "central obesity is an issue in acute myocardial infarction", there is a risk ratio of 1.1 that the central obesity of the patient concerned will be an issue in the probability of acute myocardial infarction.

Data integration of this form remains scalable over large numbers of both risk factors and users, since each observable condition is only ever evaluated with respect to *one* patient at a time, and, for clinical relevance, only ever with regard to (a small set of) that patient's most recent measurements.

7 Evaluation

To test the expressive utility of the risk ontology, as well as to populate it with data for use with QS data, a group of 8 medical doctors (members of the CARRE project team) reviewed current medical literature to identify major risk factors related to cardiorenal syndrome. At this time, 96 different risk factors were identified and described formally. The evidence sources used were 60 scientific publications. The evidence selection methodology and the available descriptions in text (tabular) format are provided in CARRE Deliverable 2.2 available from the project site [26]. A web entry system [27] allows these descriptions to be entered and reviewed, and produces RDF data representing their contents in accordance with the ontology. The manual curation of this data is necessary for regulatory and ethical reasons: as the aim of the system is to be used with patients, it is important to maintain strict quality control.

In addition, 10 project members connected a range of QS devices to the data aggregators and used or wore them to build up a sample corpus of semantically-annotated QS data. The aggregators collected data over a period of at least 12 months for all users (some users wore devices for longer), and stored them as RDF (with an average of 110,483 triples per user, at the time of writing). This length of time allowed the overall physical activity patterns of each user to be determined at different times of year and in different conditions, and thoroughly tested the data aggregators, and, importantly, was able to capture measurements which vary slowly over time, such as body weight. Other measurements, such as blood pressure, do not typically need to be measured over a long period of time to be useful in risk calculations – although it is worth noting that this commonly-held belief may simply result from a lack of data, as the ability to capture such measurements over long periods easily from non-hospitalized subjects is a comparatively recent development.

The rules expression evaluator, which evaluates the observable conditions and calculates a risk ratio for an individual for the target of a particular risk factor, is applied

to each user, for each risk association stored in the system. The same calculations were performed manually to check the fidelity of the knowledge capture. (Despite the quantity of potential risk factors, this manual process can be streamlined effectively by discarding all those risk factors which can *never* apply to a particular user – e.g., those which only apply to male populations need never be evaluated for female users.)

The risk ontology population process resulted in 253 respective associations from 96 risk factors. There were 53 involved risk elements, corresponding to a total of 90 different observables. This is an order of magnitude greater than the observables taken account of by existing risk calculators. The automatic calculation of risks agreed with the manual calculation in *every* case. It should be noted that, of course, this assesses solely whether the risk calculations are faithful to the evidence sources, not whether the evidence source itself provides good predictions (already validated via the original clinical systematic review processes) nor what, if any, effect our approach has on user behaviour to minimise risk; this will be the subject of an upcoming randomized controlled trial. For reasons of confidentiality, particularly given the small and potentially deanonymisable set of participants, the QS data cannot be made public. The online tools, however, permit reproducibility testing with new participants.

This process of testing and using the risk ontology resulted in the following qualitative findings, derived via a focus group analysis of the testing participants. The medical experts found the model straightforward to use to describe risk factors. The terminology used was found to be familiar and thus easy to understand and apply to describe risk factors found in the literature and also to read descriptions already produced by colleagues. The only difficulty identified related to expressing accurately and rigorously the observable condition that has to be satisfied in order for a risk association to hold. Initially, medical experts were asked to produce this condition in the conventional way it is written in the literature, using natural language – which was a straightforward task. Subsequently, they were asked to reformat this condition using a logical operator expression (so that this expression can be easily translated to computer readable format). This task proved to be more cumbersome and required 1-2 hours training and testing before the medical experts could independently produce correct expressions.

By using standard semantic technologies, it is possible to link both model and data to other clinical models (such as OCRe and OBX trial and data descriptions) and to external sources of data (e.g., environmental risk factors could be linked to open sources of environmental data). In particular, the semantic annotations on observables relating to medical diagnoses have made it possible to integrate the QS aggregation with Personal Health Record systems, by using UMLS to identify relevant medical concepts. Because of the semantic nature of the model, the outputs of risk calculation are also more useful for automated analysis, since it is always clear what a risk ratio value *means* in probability terms.

Nothing in either model is specific to the motivating domain of cardiorenal conditions, and extension to risk factors relating to other domains of medicine is not anticipated to pose any problems; the terminology and working practices with regard to risk calculation are common across medicine. Extending to more ‘distant’ domains where

evidence-based risk calculation is relevant (e.g., climate science) ought also to be practical. The ontology already accommodates different representations of probability, and so could be adapted to those representations suitable to the new domain's conventions. The concept of "observable" is already generic. It would be necessary to extend the notion of evidence, and in particular, evidence quality, which is currently dependent on medical definitions.

The measurement ontology has also proved to be reusable. Having been conceived as a model for capturing numerical time series data from QS devices, it has proved to be conveniently usable without modification to represent qualitative data, such as that relating to diagnoses and the severity of conditions, as well as, in preliminary work, data relating to changes in patient state. For example, if a patient becomes higher risk for a particular outcome, it is proving to be both natural and useful to clinicians to record the state change as an observation of the patient.

While the motivation and initial thinking was focused on factors which increase the probability of negative consequences, the end result is equally as capable of modelling factors which *decrease* those probabilities, or which increase the probability of *positive* consequences. In other words, it is just as straightforward to represent, for example, an intervention with the potential to *lower* a patient's chance of acute myocardial infarction as a risk association with a risk ratio less than 1. It is interesting to note that this flexibility came as something of a surprise to the medical experts on the project – it appears that the linguistic conventions in medical practice around terms such as "risk factor" and "effectiveness of treatment" obscure, to some degree, the common probabilistic structure underneath – and required a shift in philosophical approach from the clinicians to accommodate. In the same way, having to make explicit the observable conditions for grounding risk predictions in data also required a change in thinking, where conditions easily understood by experienced humans need to be spelled out in precise detail in order to be implementable. Both of these changes in thinking were seen as positive by the clinicians involved. While only a qualitative observation of a small number of people, it is perhaps reasonable to expect similar changes in thinking to be necessary for domain experts in other fields where Semantic Web approaches become more practical and applicable to more situations, and it suggests an interesting avenue for future research into the social aspects of the move to data-based approaches.

Another benefit of modelling risks explicitly in this way is that it gives a very easy to follow overview of the field of medicine under consideration, showing at a glance both which risks are increased by multiple factors, which factors lead to multiple risks, as well as which associations have received more (or less) research attention. **Fig. 6** illustrates a projection of the various risk factors, as captured by the medical experts in the context of our project. Highlighted is the example of age and ischemic heart disease increasing a patient's risk of a stroke. It can also be seen how many risk elements increase the risk of heart failure, and how many new risks appear in obese patients. Again, this is suggestive of an interesting avenue for future research, to see what may be discovered by analysis of the semantic risk data as a whole with regard to the medical research field of which it represents the output. The semantic nature of our representation is likely to be a significant advantage in such research, enabling, as

it does, the integration of the wide variety of different data sources which can be relevant to the study of scientific endeavour.

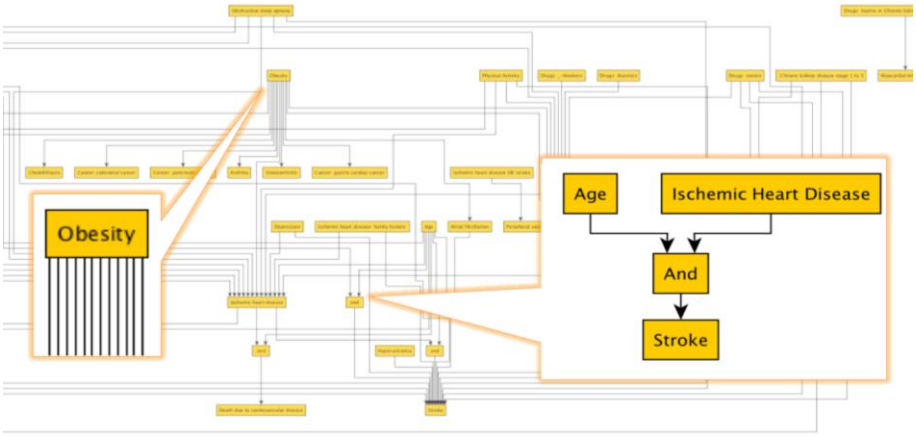

Fig. 6. A visual overview of currently encoded risk factors, with some examples highlighted, available online at http://ontology.carre-project.eu/.

8 Conclusion

The risk model presented in this paper enables clinical experts to encode the risk associations between biological, demographic, lifestyle and environmental elements and clinical outcomes in accordance with evidence from the clinical literature. The measurements model enables the automatic capture of Quantified Self data relating to individual patients in a semantically annotated form. The integration of these datasets by means of the "observable condition" rule language makes it possible to compute risks automatically.

Compared to existing risk prediction models, this approach has a significant advantage in being able to be expanded and updated easily as clinical knowledge increases and changes, as well as being transparent and traceable in function and origin. The Semantic Web approach simplifies and encourages the integration of both clinical knowledge and QS data with other sources of relevant data, and, crucially, allows an area of very complex meanings to be expressed in a machine-readable fashion. We have also shown unanticipated extra benefits of having explicit ontological models relating these types of data. In particular, analysis of risk data en masse may provide insight into the current state of overall knowledge regarding a clinical domain, and the process of knowledge capture with clinical experts required some interesting, and positive, changes in thinking and approach, drawing out commonalities and possibilities which had not before been seen. We argue that such insights are likely to be encountered in other complex domains to which Semantic Web techniques are applied.

The work presented here illustrates the value of applying the Semantic Web to Quantified Self and health data, both in and of itself and also as an illustration of us-

ing semantics to connect sources of data at very different levels of granularity and acquired through very different methods. The development of the rules language was vital to enabling our results, and we believe it would be beneficial to explore the general question of the use of rules to "bridge" distinct data sources in this way.

Acknowledgments

This work was supported by the FP7-ICT project CARRE (Grant No. 611140), funded in part by the European Commission. We express our gratitude to all project team members for fruitful discussions.

References

1. Sheridan, S., Pignone, M., & Mulrow, C. (2003). Framingham-based tools to calculate the global risk of coronary heart disease. *Journal of general internal medicine*, *18*(12), 1039-1052
2. Boon, N., Boyle, R., Bradbury, K., Buckley, J., Connolly, S., Craig, S., ... & Wood, D. (2014). Joint British Societies' consensus recommendations for the prevention of cardiovascular disease (JBS3). *Heart*, *100*(Suppl 2), ii1-ii67.
3. Woodward, M., Brindle, P., & Tunstall-Pedoe, H. (2007). Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart*, *93*(2), 172-176.
4. Kong YM, Dahlke C, Xiang Q, Qian Y, Karp D, Scheuermann RH, Toward an ontology-based framework for clinical research databases, J. Biomed. Inform., 44(1):48-58, 2011.
5. CDISC, http://cdisc.org, Accessed on: 24/07/2015.
6. Sim, I., Tu, S. W., Carini, S., Lehmann, H. P., Pollock, B. H., Peleg, M., & Wittkowski, K. M. (2014). The Ontology of Clinical Research (OCRe): an informatics foundation for the science of clinical research. *Journal of biomedical informatics*, *52*, 78-91.
7. Third, A., Kaldoudi, E., Gkotsis, G., Roumeliotis, S., , K. and Domingue, J. (2015) Capturing Scientific Knowledge on Medical Risk Factors, Workshop: 1st International Workshop on Capturing Scientific Knowledge at 8th International Conference on Knowledge Capture, Palisades, NY, USA.
8. Rivera Pelayo, V. (2015). Design and Application of Quantified Self Approaches for Reflective Learning in the Workplace, KIT Scientific Publishing, Karlsruhe.
9. Cena, F., Likavec, S., Rapp, A., Deplano, M., & Marcengo, A. (2014). Ontologies for Quantified Self: a Semantic Approach. In *HT (Doctoral Consortium/Late-breaking Results/Workshops)*.
10. Semantic Web Rules Language, http://www.w3.org/Submission/SWRL/
11. Rules Interchange Format, https://www.w3.org/TR/rif-overview/
12. Reasoners and Rule Engines: Jena Inference Support, https://jena.apache.org/documentation/inference/
13. RIF RDF and OWL compatibility, http://www.w3.org/TR/rif-rdf-owl/
14. Mrazek, P. B., & Haggerty, R. J. (Eds.) (1994). Reducing risks for mental disorders: Frontiers for preventive intervention research: Summary. National Academies Press.
15. Kaplan S, The words of risk analysis, Risk Analysis, 17(4):407-417, 1997.
16. Offord DR, Kraemer HC, Risk factors and Prevention, EBMH vol 3, p. 71, 2000.

17. Crowson CS, Therneau TM, Matteson EL, Gabriel SE, Primer: demystifying risk - understanding and communicating medical risks. Nature Clinical Practice Rheumatology, 3(3, March 2007), 2007.
18. ICD-10: International Classification of Diseases v10, WHO, http://www.who.int/classifications/icd/en/
19. SNOMED-CT: Systemized Nomenclature of Medicine – Clinical Terms, IHTSDO, http://www.ihtsdo.org/snomed-ct/
20. RxNorm: Normalized Names for Clinical Drugs, U.S. National Library of Medicine http://www.nlm.nih.gov/research/umls/rxnorm/
21. EnvO: Environmental Ontology, http://environmentontology.org/
22. QUDT: Quantity, Unit, Dimension and Type Ontologies, http://qudt.org/
23. UO: The Ontology of Units of Measurement, OBO Foundry Initiative, https://code.google.com/p/unit-ontology/
24. Oxford Centre for Evidence-based Medicine Levels of Evidence (2011) Produced by J. Howick, I. Chalmers, P. Glasziou, T. Greenhalgh, C. Heneghan, A. Liberati, I. Moschetti, B. Phillips, H. Thornton, O. Goddard and M. Hodgkinson.
25. UMLS: The Unified Medical Language System, US National Library of Medicine, http://www.nlm.nih.gov/research/umls/
26. CARRE, http://carre-project.eu/
27. CARRE Risk Data Entry system, https://entry.carre-project.eu/
28. FOAF, http://xmlns.com/foaf/spec/
29. LOINC, https://loinc.org
30. CMO, https://bioportal.bioontology.org/ontologies/CMO
31. Virtuoso Universal Server, http://virtuoso.openlinksw.com
32. Fitbit, https://www.fitbit.com
33. Medisana, http://www.medisana.com
34. iHealth, http://www.ihealthlabs.com
35. Withings, http://www.withings.com